

PATENT APPLICATION

Disk Array System Capable of Taking Over Volumes Between Controllers

Inventors: **Yoichi Mizuno**
Residence: Yokohama, Japan
Citizenship: Japan

Naoto Matsunami
Residence: Hayama-machi, Japan
Citizenship: Japan

Yasuyuki Mimatsu
Residence: Yokohama, Japan
Citizenship: Japan

Kenichi Takamoto
Residence: Odawara, Japan
Citizenship: Japan

Assignee: **Hitachi, Ltd.**
6, Kanda Surugadai 4-chome
Chiyoda-ku, Tokyo, Japan
Incorporation: Japan

Entity: Large

- 1 -

DISK ARRAY SYSTEM CAPABLE OF TAKING OVER
VOLUMES BETWEEN CONTROLLERS

BACKGROUND OF THE INVENTION

The present invention relates to a disk array system and a take-over method for logical units between controllers, and more particularly to a disk array
5 system each of whose controllers has a dedicated cache and a take-over method for logical units between controllers.

A conventional disk array system has a plurality of disks and controllers and a shared cache
10 accessible by each controller.

Fig. 10 is a block diagram showing the structure of a disk array system having a shared cache according to conventional techniques. Conventional techniques will be described with reference to Fig. 10.
15 In Fig. 10, reference numeral 100 represents a large disk array apparatus, reference symbol 200x (x = a, ..., n) represents controllers, reference numeral 300 represents an expensive shared cache, reference numeral 400 represents a shared memory, reference
20 symbol 500x (x = a, ..., n) represents disk apparatuses, and reference numeral 600 represents a common bus.

The disk array apparatus 100 shown in Fig. 10 constitutes a disk array system together with other

W0090-30/14

5

20

the management of the disk apparatus 500x. With the conventional disk array system having the shared cache, by changing the disk apparatus under the management by some controller, the access amount to logical units of
5 the disc apparatus under the management by the controller can be uniformalized so that load distribution by controllers is possible.

With the above-described conventional techniques, a disk may be mounted or dismounted to
10 change the physical position of the disk to realize load distribution. In this case, the controller stops an access to the subject disk and completely disconnects the disk from the system, and when the disconnected disk is moved to a physically different
15 position, the disk is again recognized to resume the process. With the above-described conventional techniques, a disk may be moved to a different disk apparatus. Therefore, the re-recognized disk is assigned to another new controller to allow load
20 distribution.

SUMMARY OF THE INVENTION

If the above-described conventional techniques are applied to a disk array apparatus having a shared cache, the disk array apparatus operates
25 finely. However, the conventional techniques cannot be applied to a disk array apparatus of the type that each controller has its own cache.

In a disk array system in the distributed cache environment wherein each controller has an independent and dedicated cache, each controller manages each volume by using the dedicated cache.

- 5 Therefore, access concentration upon the shared cache environment can be avoided and cost performance and scalability can be improved.

However, in the disk array system in the distributed cache environment wherein each controller
10 has an independent and dedicated cache, even if a disk apparatus under the management of one controller is to be switched to another controller, this switching is impossible because the destination controller has input/output data of volumes to be switched.

- 15 This problem can be solved if the same contents are written in the data caches of all controllers, and the take-over of volumes becomes possible similar to the shared cache. However, data of volumes not managed by each controller is also written
20 in each cache. Therefore, the cache capacity is lowered and the merit of distributed caches is lost.

The disk array system of the type that each controller has an independent and dedicated cache, is associated with the problem that automatic and dynamic
25 switching of disks cannot be performed. Namely, the disk array system in the distributed cache environment requires a manual work of disconnecting a disk from the system, a manual work of changing the physical position

1007999 03490

of the disconnected disk and a manual work of re-
recognizing the disk. During these manual works, an
access to the disk is required to be completely
stopped. Moving the disk to the original position has
5 not been taken into consideration.

As described above, the conventional disk
array system of the type that each controller has an
independent and dedicated cache is associated with the
problem that a current in-charge controller cannot be
10 replaced dynamically unless the system is stopped.

It is an object of the present invention to
solve the above-described problems and provide a disk
array system of the type that each controller has an
independent and dedicated cache and takes charge of
15 respective volumes, the system allowing any controller
to dynamically take over a desired volume of another
controller, and a method and apparatus for taking over
logical units between controllers.

According to one aspect of the present
20 invention, there is provided a disk array system having
at least one or a plurality of computers and a disk
array apparatus to be used by the computer, the disk
array apparatus having a plurality of disk apparatuses
and a plurality of controllers each having a dedicated
25 cache, wherein: the controller comprises configuration
management means for managing configuration information
of logical units formed for the disk apparatus, the
configuration management means rewrites the

1007999 "031300
2007999 "031300

configuration information of the logical units in accordance with control information to perform a take-over process of switching a desired logical unit or units under control of one controller to another
5 desired controller or controllers.

According to another aspect of the present invention, there is provided a take-over method for logical units between controllers in a disk array system having at least one or a plurality of computers
10 and a disk array apparatus to be used by the computer, the disk array apparatus having a plurality of disk apparatuses and a plurality of controllers each having a dedicated cache, the method comprises the steps of: making the controller manage configuration information
15 of logical units formed for the disk apparatus; and by rewriting the configuration information of the logical units in accordance with control information, performing a take-over process of switching a desired logical unit or units under control of one controller
20 to another desired controller or controllers.

Other objects, features and advantages of the present invention will become apparent from the following description of the embodiments of the invention taken in conjunction with the accompanying
25 drawings.

Other objects, features and advantages of the invention will become apparent from the following description of the embodiments of the invention taken

in conjunction with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram showing the configuration of a disk array system according to a first embodiment of the invention.

Fig. 2 is a diagram showing an example of the structure of a configuration information table possessed by configuration management means of the embodiment shown in Fig. 1.

Fig. 3 is a flow chart illustrating an LU take-over process between controllers according to the embodiment.

Figs. 4A and 4B are diagrams showing a specific example of the LU take-over described with reference to Fig. 3.

Figs. 5A and 5B are diagrams showing an example of the configuration information tables before and after take-over according to the embodiment.

Fig. 6 is a block diagram showing the configuration of a disk array system according to a second embodiment of the invention.

Fig. 7 is a flow chart illustrating an LU take-over process between controllers by using a data mirroring function according to the second embodiment.

Figs. 8A and 8B are diagrams showing a specific example of the LU take-over described with reference to Fig. 7.

Figs. 9A and 9B are diagrams showing an example of the configuration information tables before and after take-over according to the second embodiment.

Fig. 10 is a block diagram showing the configuration of a disk array system with a shared cache according to the prior art.

DETAILED DESCRIPTION OF THE EMBODIMENTS

Embodiments of a disk array apparatus according to the invention will be described in detail with reference to the accompanying drawings.

Fig. 1 is a block diagram showing the configuration of a disk array system according to the first embodiment of the invention, and Fig. 2 is a diagram showing an example of the structure of a configuration information table possessed by configuration management means. In Figs. 1 and 2, reference numeral 1 represents a disk array apparatus, reference symbol 2x (x = a, b, ..., n) represents computers, reference numeral 3 represents a path controller, reference numeral 4 represents a management console, reference symbol 5x (x = a, b, ..., n) represents channel paths, reference symbol 6x (x = a, b, ..., n) represents local area networks (LAN), reference numeral 7 represents communication means, reference numeral 10 represents a device network, reference symbol 11x (x = a, b, ..., n) represents controllers, reference symbol 12x (x = a, b, ..., n)

represents disk apparatuses each including general disks of, for example, 144 GB, reference numeral 31 represents path change means, reference numeral 41 represents a management utility, reference symbol 111x (x = a, b, ..., n) represents data caches, reference symbol 112x (x = a, b, ..., n) represents configuration management means, and reference symbols 121a to 121f represent logical units (LU).

As shown in Fig. 1, in the first embodiment of the invention, the disk array apparatus 1, a plurality of computers 2x using the disk array apparatus 1, the path controller 3 for interconnecting all the computers 2x, the disk array apparatus 1 and management console 4, and the management console 4 for managing the disk array apparatus 1 are interconnected by the channel paths 5x. In the example shown in Fig. 1, fiber channels are used as the channel paths. This embodiment provides LAN 6x for communications between the computer 2x and management console 4 and the communication means 7 for communications between the disk array apparatus 1 and management console 4.

The path controller 3 has the path change means which is a program for changing an access path from the computer 2x to disk array apparatus 1. In the example shown in Fig. 1, although the path change means 31 is installed in the path controller 3, a similar function may be installed in the computer 2x.

The management console 4 has the management

utility 41 which is used by a system administrator to display the LU configuration of LU's in the disk apparatus 12x of the disk array apparatus 1, to set and diagnose LU's in the disk array apparatus 1 and to perform other operations. Although the management utility 41 having a function of designating an LU take-over destination controller or switch-destination controller is installed in the management console 4, it may be installed in the computers 2x and disk array apparatus 1. Control information for designating the destination controller may be generated by the disk array apparatus.

The disk array apparatus 1 has a plurality of controllers 11x and a plurality of disk apparatuses 12x interconnected by the device network 10. With this arrangement, any controller can access any disk apparatus. Although the device network 10 may be configured by any interface, it is preferable to use fiber channels excellent in connectivity. In order to improve the total throughput, the device network may be configured by one or a plurality of switches.

The controller 11x has the data cache 111x and configuration management means 112x. The data cache temporarily stores input/output data including data written by the computer 2 and data read from the disk apparatus 12x. The configuration management means 112x manages the configuration of disks and LU's. LU (logical unit) is a virtual logical volume set in the

disk array apparatus 1. This name LU is defined by the small computer system interface (SCSI) which is one of the protocols for interface between the computer 2 and disk array apparatus. The number for identifying LU is
5 called a logical unit number (LUN) or logical volume number.

LU defined in the disk array apparatus 1 is called an internal logical volume (internal LU). In order to manage the internal LU, the disk array
10 apparatus 1 affixes integer serial numbers starting from 0 to LU's. This serial number is called an internal logical volume number (internal LUN). In detecting LU, the computer 2x sequentially searches LUN's starting from 0, and if the target number does
15 not exist, the computer stops the search of the following serial numbers in some cases. In this case, the internal LUN is not assigned to the computer 2x, but it is necessary to convert the internal LUN into LUN capable of recognized by the computer 2x. LUN
20 capable of recognized by each computer 2x is called an external logical volume (external LUN) to distinguish from the internal LUN.

The logical volumes LU 121x (121a, 121b, ..., 121n) are generated on the respective disk apparatuses
25 12x. The logical volumes LU 121x may use any RAID structure. When the logical volume LU is generated, the controller taking charge of LU is assigned. A plurality of LU's may be assigned to one controller.

In the disk array system having the configuration shown in Fig. 1, each controller 11x has a dedicated data cache 111x and takes charge of respective LU's. It is therefore possible to prevent concentration of accesses upon a particular data cache. As compared to the system using a shared cache, cost performance and scalability can be improved.

Fig. 2 shows an example of a configuration information table 1121x possessed by the configuration management means 112x. The configuration information table 1121x is used for managing the configuration information of each LU 121x. This table stores internal LUN's, external LUN's, port numbers, controller numbers, block numbers, RAID group numbers, RAID levels, physical address information and the like. This configuration information table 1121x may be stored at a physically different position, i.e., in the disk apparatus 12x at a predetermined position.

The port number is the identification number of the port usable by LU, among fiber channel connection ports of the controller 11x. The controller number is the identification number of each controller in the disk array apparatus 1. A default controller number indicates the original controller which should take charge of LU. A current controller number indicates the controller which currently takes charge of LU. The block number indicates the number of logical blocks of LU. The size of each LU can

therefore be known from the block number.

The disk array apparatus 1 may manage therein LU's of the same RAID level. The RAID group number and RAID level are the identification number of the RAID group and its level. The physical address information is physical disk position information corresponding to the logical address of each LU.

Fig. 3 is a flow chart illustrating a take-over process for LU between controllers. The take-over process for LU between controllers will be described.

(1) An operator of the management console 4 designates LUN's of all LU's to be switched and a switch-destination controller. In this case, if the port number and external LUN to be used by the switch-destination controller is to be changed, the new port number and external LUN are also designated. (Step 801)

(2) The management utility 4 issues an LU take-over instruction to the disk array apparatus 1 via the communication means 7 in accordance with the designated information. (Step 802)

(3) Upon reception of the LU take-over instruction, the configuration management means 112x of the switch-source controller 11x migrates the configuration information of the subject LU to the switch-destination controller 11x. More specifically, the configuration management means 112x of the switch-source controller 11x writes the field of the subject LU in the configuration information table 1121x into

the corresponding disk at a predetermined position, and the configuration management means 112x of the switch-destination controller 11x reads the written field to migrate the configuration information to the switch-destination controller 11x. If the port number and external LUN are to be changed, the configuration management means 112x of the switch-source controller 11x changes the contents of the configuration information table 1121x and writes the changed contents in the disk. Migration of the configuration information is performed by transmitting it via the device network 10. The configuration information may be directly transmitted via an unrepresented private line. It is obvious that the configuration management means 112x of all the controllers 11x may store in advance the configuration information of all LU's in the configuration information tables 1121x. In such a case, migration of the configuration information can be omitted, and only the port number, external LUN and controller number are changed. (Step 803)

(4) Next, the switch-source controller 11x de-stages all data of the subject LU loaded in the data cache 111x. The de-stage process is a process of writing data on the data cache into the disk. During this process, all write accesses from the computer 2x are made write-through. In this manner, all data of the subject LU can be removed from the data cache and the coherence of the contents of LU on the disk can be

maintained. The de-stage process may be performed in response to a write command from the management console 4. (Step 804)

(5) After the de-stage process for the subject LU is completed, the switch-source controller 11x notifies a de-stage completion to the management utility 41. Upon reception of the de-stage completion, the management utility 41 instructs the pass change means 31 to switch a path. (Step 805).

(6) Upon reception of this path switch instruction, the path change means 31 temporarily suspends an access to the subject LU in order to suppress an access in the transient period during the LU take-over. (Step 806)

(7) Next, the path change means 31 changes an access path to the subject LU from the switch-source controller to the switch-destination controller. All frames to the subject LU are therefore transmitted to the switch-destination controller. (Step 807)

(8) Next, the controller for controlling the subject LU is switched to the switch-destination controller. The switch-source controller completely stops an access to the subject LU, whereas the switch-destination controller starts controlling the subject LU. (Step 808)

(9) After the switching of the control is completed, the path change means 31 resumes an access to the subject LU. (Step 809)

With these Steps, the LU take-over process is completed. With these Steps, the data of the subject LU on the data cache is de-staged to maintain the coherence of the contents of the disk. Even if each
5 controller has a dedicated cache, take-over of volumes can be performed between desired controllers without suspending the system.

Figs. 4A and 4B are diagrams showing a specific example of the LU take-over described with
10 reference to Fig. 3. With reference to Figs. 4A and 4B and Figs. 5A and 5B, a specific example of the LU take-over and rewriting the configuration information table will be described.

It is assumed as shown in Fig. 4A that some
15 computer 2x uses LU0 121a to LU3 121d assigned to the controller 0 11a. It is also assumed as shown in Fig. 4B that the controller 1 11b is added to switch the control of LU2 121c and LU3 121d to the controller 1 11b. It is assumed in this case that the external LUN
20 is not changed but only the port number is changed. It is also assumed for the purpose of simplicity that all of the configuration management means 112x store the information of all LU's in the configuration information tables 1121x.

25 The configuration information tables 1121x before take-over of the switch-source and switch-destination controllers have the contents shown in Fig. 5A. The block number, RAID group number, RAID level,

physical address information and the like are omitted in Figs. 5A and 5B because they are not altered by take-over.

It can be understood from Fig. 5A that "0" representative of a current in-charge controller is stored as the default controller numbers and current controller numbers for all LU's and that LU0 and LU1 use the port number 0 and LU2 and LU3 use the port number 1.

10 An operator of the management console 4 selects LU2 121c and LU3 121d to be migrated and designates the port number "0" to be used by the switch-source controller 1 11b and switch-destination controller 1 11b. In accordance with the designated
15 information, the contents of the configuration information tables 1121x are altered. Therefore, the contents of the configuration information tables 1121x become as shown in Fig. 5B. It can be understood from Fig. 5B that the port numbers of LU2 and LU3 are
20 changed to "0" and the default controller number and current controller number are changed to "1". The switch-source controller and switch-destination controller refer to the configuration information tables and realize the logical unit configuration after
25 a change in the contents. In the example of this embodiment, control of LU2, 121c and LU3 121d can be changed to the controller 1 11b.

According to the first embodiment of the

invention, data of the subject LU on the data cache is de-staged to the disk to maintain the coherence of the contents of the disk. Therefore, even if each controller has a dedicated cache, take-over of volumes
5 can be performed between desired controllers without suspending the system, and the load distribution of each controller in the system can be realized.

Fig. 6 is a block diagram showing the configuration of a disk array system according to the
10 second embodiment of the invention. In Fig. 6, reference numeral 20 represents a tape read/write apparatus, and other reference numerals are the same as those shown in Fig. 1.

Different points of the second embodiment of
15 the invention shown in Fig. 6 from the first embodiment shown in Fig. 1 reside in that the computer 2n is connected to the tape read/write apparatus 20 and each computer 2x is directly connected to a corresponding controller 11x. Similar to the first embodiment, it is
20 obvious that the computer 2x and controller 11x may be interconnected via a path controller. The second embodiment can omit the path change means. It is not necessary that the computer 2x and controller 11x are in one-to-one correspondence, but a plurality of
25 computers may be connected to one controller.

The second embodiment of the invention shown in Fig. 6 has a data mirroring function and the computer 2n is used as a computer generally called a

backup server for backing up the volumes in the disk
array apparatus 1. The second embodiment of the
invention is applicable not only to the backup process,
but also to various processes such as batch processing
5 and data mining. In the following description, it is
assumed that the invention is applied to the backup
process.

In the disk array system shown in Fig. 6, it
is assumed that the controller 11a is assigned LU0 121g
10 and LU1 121h and the controller 11b is assigned LU2
121i and LU3 121j. LU1 121h and LU3 121j are
duplicates of LU0 121g and LU2 121i, respectively, and
store the same contents as the data stored in LU0 121g
and LU2 121i. A function of forming a duplicate of a
15 volume in the disk array apparatus is generally called
a data mirroring function.

Fig. 7 is a flow chart illustrating an LU
take-over process for LU between controllers by
utilizing the data mirroring function. This take-over
20 process will be described. As an example of the take-
over process, the case will be described wherein the
computer 2n as the backup server backs up the contents
of the duplicated LU into the tape read/write apparatus
20.

25 (1) First, the computer 2x sends to the disk
array apparatus 1 a pair-split command for LU to be
backed up. (Step 901)

(2) After the pair is split, the

configuration management means 112x of the switch-source controller 11x in charge of the split LU's migrates the configuration information of the subject LU to the controller 11n designated by the default
5 controller number in the configuration information table 1121x. For the migration, the default controller number is changed to the number of the switch-source controller. In addition to the configuration information, other necessary management information is
10 also migrated, such as the status of LU to be used by the data mirroring function, difference information for writing LU write positions and the like. The migration method may be any method similar to the first embodiment. (Step 902)

15 (3) Next, the switch-source controller 11x de-stages all the data of the subject LU on the data cache 111x. During this de-stage process, all write accesses from the computer 2x are made write-through, and the data of the subject LU on the data cache 111x
20 is removed. In this case, the computer 2x may issue a de-stage command for the subject LU. (Step 903)

(4) After the de-stage process is completed, the switch-source controller 11x notifies a de-stage completion to the switch-destination controller to
25 change control of the subject LU to the switch-destination controller 11n. (Step 904)

(5) As the control of the subject LU is changed and it becomes possible for the switch-

destination controller 11n to access the subject LU,
the computer 2n stores data of the subject LU into the
tape read/write apparatus 20 or the like to back up the
subject LU. (Step 905)

5 (6) After the backup is completed, the
backup computer 2n sends a pair re-synchronization
instruction for the subject LU to the disk array
apparatus 1. (Step 906)

10 (7) Upon reception of the pair re-
synchronization instruction, similar to the process at
Step 902 the configuration management means 112n of the
controller 11n in charge of the control of the subject
LU migrates the configuration information of the
subject LU to the switch-destination controller 11x
15 designated by the default controller number in the
configuration information table 1121n. For the
migration, the default controller number is changed to
the number of the controller 11n. In addition to the
configuration information, other necessary management
20 information is also migrated, such as the status of LU
to be used by the data mirroring function, difference
information for writing LU write positions and the
like. (Step 907)

25 (8) Next, similar to the process at Step
903, the backup controller 11n in charge of the control
of the subject LU de-stages all data of the subject LU
on the data cache 111n. During this period, all write
accesses from the computer 2n are made write-through,

20250909 10:00:00

and all data of the subject LU on the data cache 11n is removed. (Step 908)

(9) After the de-stage process is completed, the backup controller 11n in charge of the control of the subject LU notifies a de-stage process completion to the switch-destination controller 11x designated by the default controller number, and the control of the subject LU is returned to the original controller 11x to terminate the take-over process. (Step 909)

10 According to the second embodiment of the invention, since the backup process is performed by a dedicated controller 11n, data can be backed up without influencing the performance of a normal operation system.

15 Figs. 8A and 8B are diagrams showing specific examples of the LU take-over described with reference to Fig. 7, and Figs. 9A and 9B show examples of the configuration information tables before and after take-over. Next, with reference to Figs. 8A and 8B and
20 Figs. 9A and 9B, a specific example of the LU take-over and rewriting the configuration information table will be described.

It is assumed as shown in Fig. 8A that the controller 0 11a is assigned LU0 121g and LU1 121h and
25 the controller 1 11b is assigned LU2 121i and LU3 121j. It is also assumed as shown in Fig. 8B that the control of LU1 121h and LU3 121j is switched to the controller n 11n. It is assumed in this case that the external

LUN and port number are not changed but they are used as they are. Similar to the first embodiment, it is assumed that all of the configuration management means 112x store the information of all LU's in the
5 configuration information tables 1121x.

The configuration information tables 1121x before take-over have the contents shown in Fig. 9A. It can be understood from Fig. 9A that "0" is stored for LU0 and LU1 and "1" is stored for LU2 and LU3,
10 indicating the controller number of a current in-charge controller. As the default controller number for LU1 and LU3, "n" representative of the switch-destination controller is stored.

It can be understood from the configuration
15 information table of Fig. 9A that the configuration management means 112a of the switch-source controller 0 11a and the configuration management means 112b of the switch-source controller 1 11b can know that the controller n 11n represented by the default controller
20 number is the switch-destination controller.

Therefore, the contents of the configuration information table after take-over become as shown in Fig. 9B. It can be understood from this table that the default controller numbers of LU1 and LU3 are changed
25 to "0" and "1", respectively. The controller in charge of the control after take-over is the controller n and the current controller numbers are changed to n.

After the backup process is completed by

changing the controllers, the controller performs migration along a direction opposite to the first migration. In this case, similar to the first migration, the configuration management means 112n of the controller n 11n can know the switch-destination controllers from the default controller numbers in the configuration information table 1121n. It is therefore possible to automatically switch the control of LU's to the original controllers. The controllers and LU assignment when the LU control is switched to the original controllers are the same as those shown in Fig. 8A. When the control of the subjects LU's is switched to the original controllers, the default controller numbers of the subject LU's in the configuration information table 1121x are changed to n. The contents of the configuration information table are the same as those shown in Fig. 9A. It can be understood from Fig. 9A that the default controller numbers of LU1 and LU3 are again changed to n. The LU take-over can be repetitively performed in the manner described above.

According to the second embodiment, the backup process is performed by using the dedicated controller so that LU's can be backed up without influencing the performance of the normal operation system. By storing the controller numbers of the switch-destination controllers in the configuration information table, it becomes possible to automatically

take over LU's and automatically return the control to original controllers to thereafter continue the normal process.

As described so far, according to the present invention, even if each controller has an independent and dedicated cache, desired LU's can be taken over switching between desired controllers without suspending the system and load distribution of each controller in the system becomes possible.

According to the present invention, automatic volume take-over and returning the control to original controllers are possible. Accordingly, the backup process can be performed by using a dedicated controller, and LU's can be backed up compatibly with the normal operation system.

It should be further understood by those skilled in the art that the foregoing description has been made on embodiments of the invention and that various changes and modifications may be made in the invention without departing from the spirit of the invention and the scope of the appended claims.